Lecture 16.

$\boxed{\text{Topic 5}}$ Unconstrained minimization.

1)

$$\min f(x)$$

Assume $f(x): \mathbb{R}^n \to \mathbb{R}$ is convex, twice differentiable

Assume that the problem is solvable, that is, an x^* exists

Denote

$$p^* = f(x^*) = \inf_x f(x)$$

- Since $f(x)$ is differentiable and convex, for x^* to be optimal it is neccessary and sufficient that

$$\nabla f(x^*) = 0$$

If this equation has analytical solution, we are done!

- Often the above equation doesnot allow an analytical solution and hence must be solved numerically, through an iterative algorithm.

This is an algorithm that computes a sequence of points

$$x^{(0)}, x^{(1)}, x^{(2)}, \dots \in \text{dom} f$$

Such that $f(x^{(k)}) \to p^*$ as $k \to \infty$.

Such an sequence is called a minimizing sequence for the original optimization problem.

The algorithm terminates when $f(x^{(k)}) - p^* \le \varepsilon$ where $\varepsilon > 0$ is some pre-specified tolerance.

+) Initial point and sublevel set.

The methods we are going to study for unconstrained minimization here require a suitable starting point $x^{(0)}$ such that:

$$\begin{cases} x^{(0)} \in \text{dom} f \\ \text{Sub level set } S = \{x \in \text{dom} f \mid f(x) \le f(x^{(0)})\} \text{ is closed.} \end{cases}$$

Examples: (i) Least-square

$$\min \|Ax-b\|_2^2 = x^T(A^TA)x - 2(A^Tb)^Tx + b^Tb$$

The optimality condition is

$$A^TAx^* = A^Tb.$$

- If A^TA is invertable, then the problem has a unique solution.
- If $A^TA \succeq 0$ but is not strictly $\succ 0$, then any solution of the above equation is optimal.
- If the optimality equation does not have a solution then the optimization problem is unbounded below ($p^* = -\infty$).

(ii) GP: $\min f(x) = \log\left(\sum_{i=1}^{m} e^{a_i^Tx+b_i}\right)$

The optimality condition is

$$\nabla f(x^*) = \frac{1}{\sum_{i=1}^{m} e^{a_i^Tx^*+b_i}} \sum_{i=1}^{m} e^{a_i^Tx^*+b_i} \cdot a_i = 0$$

This equation does not have an analytical solution in general so we must rely on an iterative algorithm to find x^*.

+) Notes: The condition on closed sublevel sets is usually hard to verify, but holds if

$\mathrm{dom} f = \mathbb{R}^n$ (whole space)

or $f(x) \to \infty$ as $x \to$ boundary of $\mathrm{dom} f$

+) Strong convexity and implications:

Recall that the condition for $f(x)$ to be convex is

$\nabla^2 f(x) \succeq 0$ (provided $f(x)$ is differentiable).

- Here we assume strong convexity: $\exists\, m > 0$ s.t.

$$\nabla^2 f(x) \succeq mI \quad \forall x \in S.$$

That is, the Hessian is strictly positive definite (on set S)

- By the Taylor's expansion and mean-value theorem:

$\forall x, y \in S$: $\exists\, z \in [x, y]$ s.t.

$$f(y) = f(x) + \nabla f(x)^T (y-x) + \frac{1}{2}(y-x)^T \nabla^2 f(z)(y-x)$$

- By strong convexity then

$$f(y) \geq f(x) + \nabla f(x)^T (y-x) + \frac{m}{2}(y-x)^T(y-x) \quad \forall x, y \in S$$

When $m = 0$, we recover the first order condition for convexity

- For $m > 0$, we obtain a better bound.

- We can obtain a bound on p^* as follows:

Minimize the RHS expression wrt y:

minimizer $\tilde{y} = x - \frac{1}{m}\nabla f(x)$ (setting the derivatives wrt y to zero)

Hence

$$f(y) \geq f(x) + \nabla f(x)^T(\tilde{y}-x) + \frac{m}{2}(\tilde{y}-x)^T(\tilde{y}-x)$$
$$= f(x) - \frac{1}{2m}\|\nabla f(x)\|_2^2 \quad \forall y \in S.$$

$$\rightarrow \quad p^* \geq f(x) - \frac{1}{2m}\|\nabla f(x)\|_2^2.$$

This inequality shows that if $\|\nabla f(x)\|_2$ is small at a point then the point is nearly optimal.

- If we know m, we can also use it as a stopping criterion.

$$f(x) - p^* \leq \frac{1}{2m} \|\nabla f(x)\|_2^2 .$$

- We can also show a bound on the optimal point:
  Let $y = x^*$

$$\rightarrow p^* = f(x^*) \geq f(x) + \nabla f(x)^T(x^* - x) + \frac{m}{2} \|x^* - x\|_2^2$$
$$\geq f(x) - \|\nabla f(x)\| \, \|x^* - x\| + \frac{m}{2} \|x^* - x\|_2^2$$
(by Cauchy-Schwatz)

but $p^* \leq f(x) \ \forall x$

$$\rightarrow -\|\nabla f(x)\| \cdot \|x^* - x\|_2 + \frac{m}{2} \|x^* - x\|_2^2 \leq 0 \quad \forall x$$

$$\rightarrow \|x - x^*\|_2 \leq \frac{2}{m} \|\nabla f(x)\|_2 \ \forall x .$$

One implication of this is the optimal point x^* is unique.

+) Upper bound on $\nabla^2 f(x)$:
Strong convexity implies that the sublevelsets contained in S, thus S is bounded.
Thus $\exists M > 0$ such that

$$\nabla^2 f(x) \preceq MI . \ \forall x \in S .$$

This implies

$$\forall x, y \in S : \ f(y) \leq f(x) + \nabla f(x)^T (y - x) + \frac{M}{2} \|y - x\|_2^2$$

which, by minimizing both sides over y, yields

$$p^* \leq f(x) - \frac{1}{2M} \|\nabla f(x)\|_2^2$$

Thus

$$f(x) - \frac{1}{2m} \|\nabla f(x)\|_2^2 \leq p^* \leq f(x) - \frac{1}{2M} \|\nabla f(x)\|_2^2$$
$$\forall x \in S .$$

+) Condition number bound:

$$mI \preceq \nabla^2 f(x) \preceq MI \quad \forall x \in S.$$

The ratio $K = \frac{M}{m}$ is an upper bound on the condition number of $\nabla^2 f(x)$.

This condition number has a strong effect on the efficiency of some common methods for unconstrained optimization.

- Usually the constants m and M are unknown, so they cannot be used directly as a practical stopping criterion.

However, they can be used as a conceptual stopping criteri[on] and used in convergence proofs for algorithms. These convergen[ce] proofs usually assume some (unknown) constants m, M, except for a special class of convex functions (self-concorda[nt]). More on this latter.

2). Descent methods.

- Consider algorithms that produce a minimizing sequence $x^{(k)}$

$$x^{(k+1)} = x^{(k)} + t^{(k)} \Delta x^{(k)}$$

where $\Delta x^{(k)}$: step or search direction (vector in $\mathbb{R}^n$)

k : iteration number

$t^{(k)} > 0$: step size or step length ($t^{(k)} = 0$ if $x^{(k)}$ is optimal)

- When focusing on one iteration, we write

$$x^{+} = x + t\Delta x.$$

+) Condition number bound:

$$mI \preceq \nabla^2 f(x) \preceq MI \quad \forall x \in S.$$

The ratio $K = \frac{M}{m}$ is an upper bound on the condition number of $\nabla^2 f(x)$.

This condition number has a strong effect on the efficiency of some common methods for unconstrained optimization.

∘ Usually the constants m and M are unknown, so they cannot be used directly as a practical stopping criterion.

However, they can be used as a conceptual stopping criterion and used in convergence proofs for algorithms. These convergence proofs usually assume some (unknown) constants m, M, except for a special class of convex functions (self-concordant). More on this latter.

cture 17:

2). Descent methods.

∘ Consider algorithms that produce a minimizing sequence $x^{(k)}$:

$$x^{(k+1)} = x^{(k)} + t^{(k)} \Delta x^{(k)}$$

where $\Delta x^{(k)}$: step or search direction (vector in $\mathbb{R}^n$)

k : iteration number

$t^{(k)} > 0$: step size or step length ($t^{(k)} = 0$ if $x^{(k)}$ is optimal)

∘ When focusing on one iteration, we write

$$x^{+} = x + t\Delta x.$$

$$f(x) - p^* \leq \frac{1}{2m} \|\nabla f(x)\|_2^2 .$$

- We can also show a bound on the optimal point:
  Let $y = x^*$

$$\rightarrow p^* = f(x^*) \geq f(x) + \nabla f(x)^T (x^* - x) + \frac{m}{2} \|x^* - x\|_2^2$$
$$\geq f(x) - \|\nabla f(x)\| \, \|x^* - x\| + \frac{m}{2} \|x^* - x\|_2^2$$
(by Cauchy-Schwartz)

but $p^* \leq f(x) \ \forall x$

$$\rightarrow -\|\nabla f(x)\| \cdot \|x^* - x\|_2 + \frac{m}{2} \|x^* - x\|_2^2 \leq 0 \quad \forall x$$

$$\rightarrow \|x - x^*\|_2 \leq \frac{2}{m} \|\nabla f(x)\|_2 \quad \forall x .$$

One implication of this is the optimal point x^* is unique.

+) Upper bound on $\nabla^2 f(x)$:
Strong convexity implies that the sublevelsets contained in S, thus S is bounded.
Thus $\exists M > 0$ such that

$$\nabla^2 f(x) \preceq MI . \quad \forall x \in S .$$

This implies

$$\forall x, y \in S : \quad f(y) \leq f(x) + \nabla f(x)^T (y - x) + \frac{M}{2} \|y - x\|_2^2$$

which, by minimizing both sides over y, yields

$$p^* \leq f(x) - \frac{1}{2M} \|\nabla f(x)\|_2^2$$

Thus

$$f(x) - \frac{1}{2m} \|\nabla f(x)\|_2^2 \leq p^* \leq f(x) - \frac{1}{2M} \|\nabla f(x)\|_2^2$$
$$\forall x \in S .$$

- A method is called a descent method if

$$f(x^{(k+1)}) < f(x^{(k)})$$

except when $x^{(k)}$ is optimal.

- From convexity of $f(x)$, we know that

$$\nabla f(x^{(k)})^T (y - x^{(k)}) \geq 0 \text{ implies } f(y) \geq f(x^{(k)})$$

since

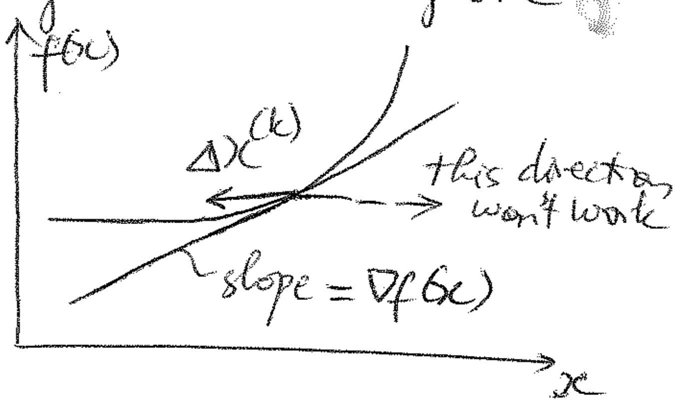
$$f(y) \geq f(x^{(k)}) + \nabla f(x^{(k)})^T (y - x^{(k)}) \quad \forall y, x^{(k)} \in \mathrm{dom} f$$

(1st order condition)

Thus the search direction in a descent method must satisfy

$$\nabla f(x^{(k)})^T \Delta x^{(k)} < 0.$$

→ it must make an acute angle with the negative gradient at $x^{(k)}$.



+) General descent method.

given a starting point $x \in \mathrm{dom} f$

repeat 1. Determine a descent direction

2. Line search : choose a step size $t > 0$

3. Update: $x := x + t\Delta x$

until stopping criterion is satisfied.

The stopping criterion is often of the form

$$\|\nabla f(x)\|_2 < \varepsilon$$

where ε is a small, positive, prechosen tolerance value

+ <u>Line search</u>: Line search is used to select the stepsize t, which determine how far along the line $x + t\Delta x$ the next iterate will be.

- <u>Exact line search</u>:

$$t = \arg\min_{s \geq 0} f(x + s\Delta x)$$

This method can be used if the minimizer of $f(x)$ along the line $x + s\Delta x$ can be found analytically, or if the cost of finding the exact minimum is low.

- Backtracking line search:

Most practical line searches are inexact: to reduce $f(x)$ "enough" long the direction Δx.

<u>Backtracking line search</u>: 2 constant parameters $\alpha \in (0, \frac{1}{2})$, $\beta \in ($

starting at $t = 1$

while $f(x + t\Delta x) > f(x) + \alpha t \nabla f(x)^T \Delta x$

$$t = \beta t.$$

(Typical values: $\alpha \in [0.01, 0.3]$, $\beta \in [0.1, 0.8]$.)

- Since Δx is a descent direction, then $\nabla f(x)^T \Delta x < 0$. Thus for t small enough we have

$$f(x + t\Delta x) \simeq f(x) + t\nabla f(x)^T \Delta x$$

$$< f(x) + \alpha t \nabla f(x)^T \Delta x$$

Hence backtracking line search will always terminate.

this comes from an approximation of quadratic as linear

[Why $\alpha < \frac{1}{2}$ will be clear in convergence analysis later].

Note: If $\mathrm{dom} f \neq \mathbb{R}^n$ then care must be taken to ensure that $x + t\Delta x \in \mathrm{dom} f$ before checking the inequality in backtracking line search.

3). Gradient descent method:

The question here is how to choose the (descent) search direction Δx?

A candidate is the negative gradient:

$$\Delta x = -\nabla f(x)$$

Gradient descent method:

given a starting point $x \in \mathrm{dom} f$.
repeat
1. $\Delta x := -\nabla f(x)$
2. if $\|\nabla f(x)\|_2 \leq \varepsilon$, break
else
3. Line search: choose t via exact or backtracking
4. Update $x := x + t\Delta x = x - t\nabla f(x)$
end.

4) Convergence analysis: We will show the convergence for backtracking line search type

Assume strong convexity holds: $\exists\, m, M > 0$ s.t.

$$mI \preceq \nabla^2 f(x) \preceq MI \quad \forall x \in S$$

Consider step length t such that $x - t\nabla f(x) \in S$.

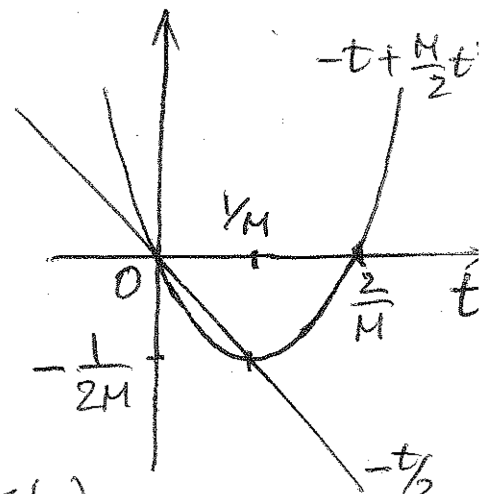
Recall $f(y) \le f(x) + \nabla f(x)^T (y-x) + \frac{M}{2}\|y-x\|_2^2 \quad \forall x, y \in S$

let $y = x - t\nabla f(x)$ then

$$f(x - t\nabla f(x)) \le f(x) - t\|\nabla f(x)\|_2^2 + \frac{M}{2}t^2\|\nabla f(x)\|_2^2 .$$

Now use the inequality

$$-t + \frac{Mt^2}{2} \le -\frac{t}{2} \quad \text{for } 0 \le t \le \frac{1}{M}$$



we have

$$\begin{aligned} f(x - t\nabla f(x)) &\le f(x) - \frac{t}{2}\|\nabla f(x)\|_2^2 \\ &\le f(x) - \alpha t\|\nabla f(x)\|_2^2 \end{aligned}$$

for $0 < \alpha < \frac{1}{2}$. (Hence the range $\alpha < \frac{1}{2}$).

Thus for t small enough $\left(0 < t \le \frac{1}{M}\right)$, the back tracking line search will terminate.

This happens either with $t = 1$ or for some $1 > t \ge \frac{\beta}{M}$.

Thus we can write

$$f(x^+) \le f(x) - \min\left\{\alpha, \frac{\beta\alpha}{M}\right\}\|\nabla f(x)\|_2^2 .$$

$$\rightarrow f(x^+) - p^* \le f(x) - p^* - \min\left\{\alpha, \frac{\beta\alpha}{M}\right\}\|\nabla f(x)\|_2^2$$

Recall also that $f(x) - p^* \le \frac{1}{2m}\|\nabla f(x)\|_2^2 \quad \forall x \in S.$

thus

$$f(x^+) - p^* \le \left(1 - 2m\cdot\min\left\{\alpha, \frac{\beta\alpha}{M}\right\}\right)(f(x) - p^*).$$

Therefore $f(x^{(k)}) - p^* \le c^k (f(x^{(0)}) - p^*)$

for a constant $c = 1 - \min\left\{2m\alpha, \frac{2\beta\alpha m}{M}\right\} < 1.$

Lecture 18:

+) Example: A quadratic problem.

$$f(x) = \frac{1}{2}(x_1^2 + \gamma x_2^2) \qquad \gamma > 0.$$

This problem has an analytical solution of $x_1 = x_2 = 0$.

$$\nabla^2 f(x) = \begin{bmatrix} 1 & 0 \\ 0 & \gamma \end{bmatrix} \rightarrow \begin{array}{l} m = \min(1, \gamma) \\ M = \max(1, \gamma) \end{array}$$

Apply gradient search with exact line search, we can obtain closed form expression for the iterates

$$x_1^{(k)} = \gamma\left(\frac{\gamma-1}{\gamma+1}\right)^k, \quad x_2^{(k)} = \left(-\frac{\gamma-1}{\gamma+1}\right)^k$$

and $f(x^{(k)}) = \left(\frac{\gamma-1}{\gamma+1}\right)^{2k} f(x^{(0)})$ for $x^{(0)} = (\gamma, 1)$.

Convergence is linear (exactly): the error reduces by the factor $[(\gamma-1)/(\gamma+1)]^2$ at each iteration.

If $\gamma = 1 \rightarrow$ algorithm takes exactly 1 iteration.

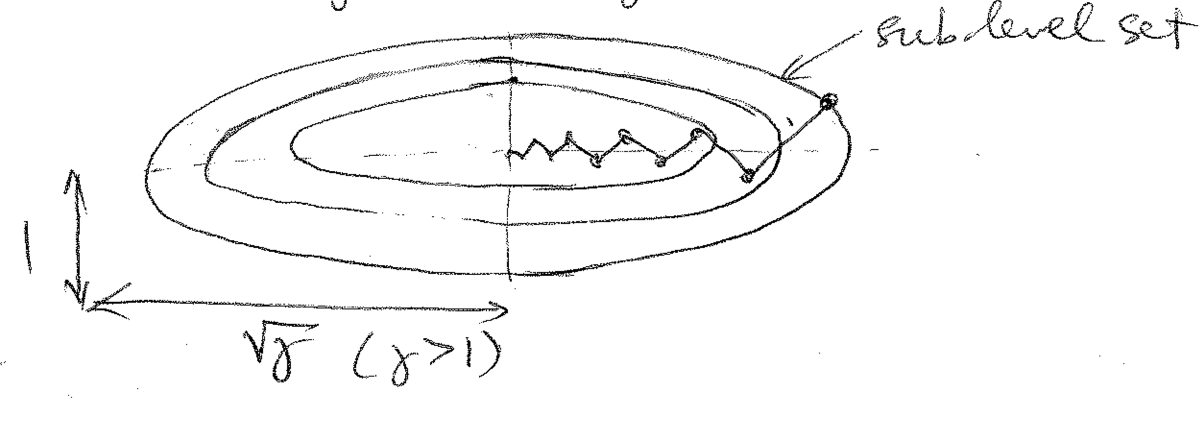
γ around 1 (say $\frac{1}{3} \leq \gamma \leq 3$) $\rightarrow$ convergence is fast.

$\gamma \gg 1$ or $\gamma \ll 1 \rightarrow$ convergence is very slow.

Using the bound derived earlier, error is reduced each iteration by exactly

$$\left(\frac{1 - m/M}{1 + m/M}\right)^2 = \frac{K-1}{K+1}$$

If $K = \frac{M}{m}$ is large $\rightarrow$ convergence is slow



+) Several observations:

- The gradient method exhibits approximately linear convergence. (error reduces by a factor of c each iteration)
- The choice of backtracking parameters α, β has noticeable but not dramatic effect on the convergence.
- The convergence rate depends strongly on the condition number of the Hessian. Convergence is slow even for $K \simeq 100$. For $K \geq 1000$, the gradient method is practically useless.

4. Steepest descent method:

This method determines the "steepest" direction according to some norm.

- First-order Taylor approximation of $f(x+v)$ around x:

$$f(x+v) \approx f(x) + \nabla f(x)^T v.$$

$\nabla f(x)^T v$: directional derivatives of f at x in direction v.

- Normalized steepest descent direction:

$$\Delta x_{nsd} = \arg\min_v \{ \nabla f(x)^T v \mid \|v\| \leq 1 \}$$

for some norm $\|v\|$.

Δx_{nsd} is the direction in the unit ball in norm $\|v\|$ that extends farthest in the direction $-\nabla f(x)$.

For the 2-norm $\|v\|_2$, steepest descent direction is the same as gradient descent.

- Unnormalized steepest descent direction:

$$\Delta x_{sd} = \|\nabla f(x)\|_* \, \Delta x_{nsd}$$

where $\|\cdot\|_*$ denotes the dual norm:

$$\nabla f(x)^T \Delta x_{sd} = \|\nabla f(x)\|_* \nabla f(x)^T \Delta x_{nsd} = -\|\nabla f(x)\|_*^2.$$

<u>Note</u>: A dual norm is given by:

$$\|z\|_* = \sup\{z^T x \mid \|x\| \leq 1\}.$$

For example: • The dual of the Euclidean norm is the Euclidean norm itself.

• The dual of the l_∞-norm is the l_1-norm.

<u>Steepest descent alg.</u>

given a starting point $x \in \mathrm{dom} f$

repeat

1. Compute the steepest descent direction Δx_{sd}.
2. Line search: choose step t
3. Update: $x := x + t\Delta x_{sd}$.

until stopping criterion is satisfied.

Convergence properties are similar to gradient descent (linear convergence)

<u>Examples</u>:

(i) Euclidean norm: $\Delta x_{sd} = -\nabla f(x)$

(ii) Quadratic norm:

$$\|z\|_P = (z^T P z)^{1/2} = \|P^{1/2} z\|_2, \quad P \in S_{++}^n$$

Steepest descent direction:

$$\Delta x_{sd} = -P^{-1} \nabla f(x)$$

Thus quadratic norm is equivalent to a change of coordinates:

Let $\tilde{u} = P^{1/2} u$ and $\tilde{f}(\tilde{u}) = f(P^{-1/2}\tilde{u}) = f(u)$

then $\nabla \tilde{f}(\tilde{x}) = -P^{-1/2} \nabla f(P^{-1/2}\tilde{x}) = P^{-1/2} \nabla f(x)$

The gradient search direction corresponds to

$$\Delta x = P^{-1/2}\left(-P^{-1/2}\nabla f(x)\right) = -P^{-1}\nabla f(x)$$

for the original variable x.

Thus steepest direction is the same as the (negative) gradient direction after the change of coordinate $\tilde{x} = P^{1/2}x$.

+) The choice of P can have a strong effect on the rate of convergence.

- If the change of coordinate by P reduces the condition number of the resulting sublevel sets $\rightarrow$ speed up convergence, and vice versa.

See figures 9.11 – 9.15 in the text for examples.

5. Newton's method:

- For gradient and steepest descent methods, we approximate the objective function at the current value $x^{(k)}$ by a linear function going through $x^{(k)}$.

- For Newton's method, we approximate it by a quadratic function through $x^{(k)}$.

+) Newton step:

$$\Delta x_{nt} = -\nabla^2 f(x)^{-1}\nabla f(x)$$

For convex functions, $\nabla^2 f(x) \succeq 0$, thus

$$\nabla f(x)^T \Delta x_{nt} = -\nabla f(x)^T \nabla^2 f(x)^{-1}\nabla f(x) < 0$$

unless $\nabla f(x) = 0$ $\rightarrow$ Δx_{nt} is a descent direction.

Interpretations:

+) Minimizer of second-order Taylor approximation:

$$\hat{f}(x+v) = f(x) + \nabla f(x)^T v + \frac{1}{2} v^T \nabla^2 f(x) v.$$

$\hat{f}(x+v)$ is a convex quadratic approximation of $f(x)$ at x.
The minimizer of $\hat{f}(x+v)$ is Δx_{nt}:

$$\frac{\partial \hat{f}}{\partial v} = v^T \nabla^2 f(x) + \nabla f(x)^T = 0$$

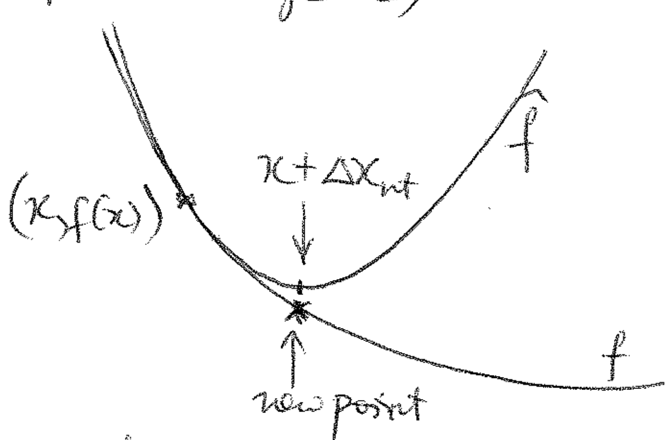
$$\longrightarrow \quad v^* = -\nabla^2 f(x)^{-1} \nabla f(x) = \Delta x_{nt}$$

Since $f(x)$ is twice differentiable, the quadratic model is a very good approximation of $f(x)$ when x is near x^*.

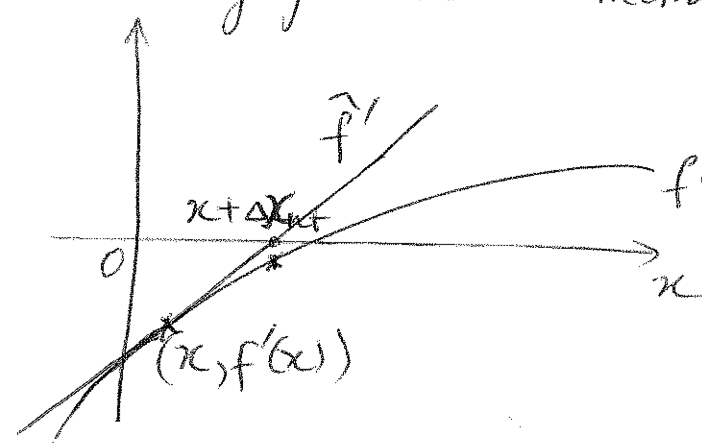
+) Steepest descent direction in Hessian norm:

$$\|u\|_{\nabla^2 f(x)} = \left(u^T \nabla^2 f(x) u\right)^{1/2}$$

When x is near x^*, we have $\nabla^2 f(x) \approx \nabla^2 f(x^*)$, which makes the Hessian after this change of coordinate to have almost a condition number of $1 \rightarrow$ very good search direction (fast convergence)



quadratic approximation.



linearized optimality cond.

+) Linearized optimality condition: $\nabla f(x^*) = 0$, linearize $\nabla f(x)$ to get

$$\nabla f(x+v) \cong \nabla f(x) + \nabla^2 f(x) v = 0 \rightarrow v = \Delta x_{nt}$$

Lecture 19:

+) Affine Invariance of the Newton step (property):

If we perform a change of coordinate: $x = Ty$, $T \in \mathbb{R}^n$
T non-singular.

Denote $\tilde{f}(y) = f(Ty) = f(x)$

$$\to \nabla \tilde{f}(y) = T^T \nabla f(Ty)$$

$$\nabla^2 \tilde{f}(y) = T^T \nabla^2 f(Ty)\, T$$

Then the Newton step for $\tilde{f}(y)$ at y is:

$$\Delta y_{nt} = -\left(T^T \nabla^2 f(x) T\right)^{-1} T^T \nabla f(x)$$

$$= -T^{-1} \nabla^2 f(x)^{-1} \nabla f(x)$$

$$= T^{-1} \Delta x_{nt}$$

Thus the Newton updates of $\tilde{f}$ and f are also related by the same affine transformation:

$$x + \Delta x_{nt} = T(y + \Delta y_{nt}).$$

+) Newton decrement:

$$\lambda(x) = \left(\nabla f(x)^T \nabla^2 f(x)^{-1} \nabla f(x)\right)^{+1/2}$$

is called the Newton decrement at x.

- It tells approximately how far point x is from the optimal.

$$f(x) - \inf_y \hat{f}(y) = f(x) - \hat{f}(x + \Delta x_{nt}) = \frac{1}{2} \lambda(x)^2.$$

$\to \frac{1}{2}\lambda(x)^2$ gives an estimate of $f(x) - p^*$ (not a lower or upper bound, just an estimate - more later).

- It gives the norm of the Newton step:

$$\lambda(x)^2 = \Delta x_{nt}^T \nabla^2 f(x)\, \Delta x_{nt}$$

- Also shows up in backtracking linesearch:

$$\nabla f(x)^T \Delta x_{nt} = -\lambda(x)^2.$$

- $\lambda(x)$ is affine invariant

$$\lambda(y) = \lambda(x) \quad \text{for} \quad \tilde{f}(y) = f(Ty) \underset{x=Ty}{=} f(x)$$

+) <u>Newton's method:</u>

given a starting point $x \in \text{dom} f$, tolerance $\varepsilon > 0$
repeat

1. Compute the Newton step and decrement
$$\Delta x_{nt} := -\nabla^2 f(x)^{-1} \nabla f(x)$$
$$\lambda^2 := \nabla f(x)^T \nabla^2 f(x)^{-1} \nabla f(x)$$
2. Check stopping criterion: quit if $\lambda^2/2 \leq \varepsilon$.
3. Linesearch: Choose step size t by backtracking linesearch.
4. Update: $x := x + t\Delta x_{nt}$.

+) <u>Convergence analysis:</u>

- Again assume strong convexity: $\exists\, m, M > 0$ s.t.

$$mI \preccurlyeq \nabla^2 f(x) \preccurlyeq MI \quad \forall x \in S.$$

- Also assume the Hessian of f is Lipschitz continuous on S with constant $L > 0$:

$$\| \nabla^2 f(x) - \nabla^2 f(y) \|_2 \leq L \|x - y\|_2 \quad \forall x, y \in S.$$

Intuitively, L represents a bound on the third derivative of f.
$L = 0$ for a quadratic function
L small $\to$ quadratic approximation of f at x is good
L large $\to$ quadratic approximation is poor.
$\Rightarrow$ L will play an important role in the convergence rate.

+) Idea of convergence proof:
Show $\exists$ η and γ:

$$0 < \eta \leq \frac{m^2}{L}, \quad \gamma > 0$$

such that there are 2 regions of convergence:

- If $\|\nabla f(x^{(k)})\|_2 \geq \eta$ then

$$f(x^{(k+1)}) - f(x^{(k)}) \leq -\gamma.$$ (damped Newton phase).

- If $\|\nabla f(x^{(k)})\|_2 < \eta$, then

$$\frac{L}{2m^2}\|\nabla f(x^{(k+1)})\|_2 \leq \left(\frac{L}{2m^2}\|\nabla f(x^{(k)})\|_2\right)^2.$$: quadratic convergen

+) Damped Newton phase:
- Function value decreases by at least γ at each iteration.
- Line search with backtracking is used
- Number of iteration in this step:

$$\#\text{ iterations} \leq \frac{f(x^{(0)}) - f^*}{\gamma}.$$

+) Quadratic convergence phase:
- Since $\eta \leq \frac{m^2}{L}$, once the algorithm enters this phase, it wil stay in this phase:

$\forall l > k$: $\|\nabla f(x^{(l)})\|_2 < \eta$ since

$$\|\nabla f(x^{(k+1)})\|_2 \leq \frac{2m^2}{L}\cdot\left(\frac{1}{2}\cdot\frac{L}{m^2}\|\nabla f(x^{(k)})\|_2\right)^2$$
$$\leq \cdot\left(\frac{1}{2}\cdot\frac{1}{\eta}\right)\cdot\eta^2 = \frac{1}{2}\eta < \eta.$$

- Apply the inequality recursively, we get:

$$\frac{L}{2m^2}\|\nabla f(x^{(l)})\|_2 \leq \left(\frac{L}{2m^2}\|\nabla f(x^{(k)})\|_2\right)^{2^{l-k}} \leq \left(\frac{1}{2}\right)^{2^{l-k}}$$

(since $\|\nabla f(x^{(k)})\|_2 < \eta \leq \frac{m^2}{L}$)

Thus

$$f(x^{(l)}) - p^* \underset{\substack{\uparrow \\ \text{strong convexity}}}{\leq} \frac{1}{2m}\|\nabla f(x^{(l)})\|_2^2 \leq \underbrace{\frac{2m^3}{L^2}}_{\varepsilon_0}\left(\frac{1}{2}\right)^{2^{l-k+1}}$$

→ Convergence is extremely rapid in this phase.

+) The overall number of iterations is bounded above by

$$\frac{f(x^{(0)}) - p^*}{\gamma} + \log_2\log_2\left(\frac{\varepsilon_0}{\varepsilon}\right)$$

$\varepsilon_0 = \frac{2m^3}{L}$

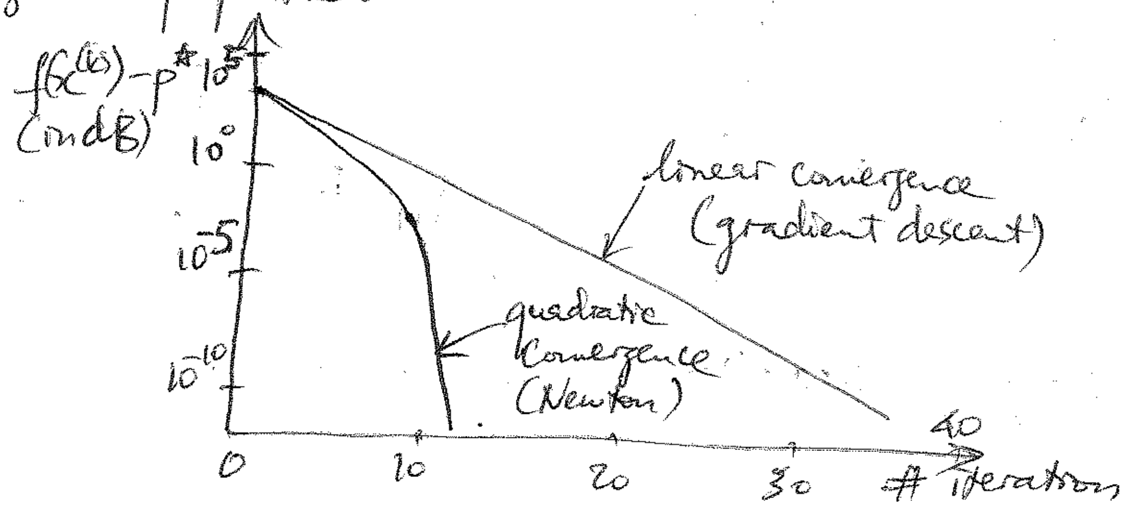
ε : prespecified tolerance level.

For example, six iterations in the quadratic phase gives an accuracy of about

$$\varepsilon \approx 5 \times 10^{-20}\,\varepsilon_0 .$$

→ the quadratic phase can be considered to have almost constant # of iteration (≈ 6).

◦ In practice, the constants m, L (hence γ, ε_0) are usually unknown. But this convergence analysis provides insight into convergence properties.



+) More detailed convergence proof:

∘ Damped Newton phase:

By strong convexity $\nabla^2 f(x) \preceq MI \quad \forall x \in S$:

$$f(x + t\Delta x_{nt}) \le f(x) + t\nabla f(x)^T \Delta x_{nt} + \frac{M}{2}\|\Delta x_{nt}\|_2^2 t^2$$

$$\le f(x) - t\lambda(x)^2 + \frac{M}{2m} t^2 \lambda(x)^2 \le f(x) - \frac{t}{2}\lambda$$

Noting step size $\hat{t} = \frac{m}{M}$ satisfies backtracking line search, then line search must return $t \ge \beta \frac{m}{M}$, thus

$$f(x^+) - f(x) \le -\frac{t}{2}\lambda(x)^2 \le -\alpha t \lambda(x)^2$$

$$\le -\frac{\alpha\beta m}{M^2}\|\nabla f(x)\|_2^2 \le -\alpha\beta\frac{m}{M^2}\eta^2$$

[Here we use $\lambda(x)^2 = \nabla f(x)^T \nabla^2 f(x)^{-1} \nabla f(x)^T \ge \frac{1}{M}\|\nabla f(x)\|_2^2$

Thus choose $\gamma = \alpha\beta\frac{m}{M^2}\eta^2$ satisfies the damped phase.

∘ Quadratic phase:

For this phase we can always use step size $t = 1$ in backtracking search. We skip the proof for this here.

Now $x^+ = x + \Delta x_{nt}$ and

$$\|\nabla f(x^+)\|_2^2 = \|\nabla f(x + \Delta x_{nt}) - \nabla f(x) - \nabla^2 f(x)\Delta x_{nt}\|_2$$

$$= \left\| \int_0^1 \nabla^2 f(x + t\Delta x_{nt}) \cdot \Delta x_{nt}\, dt - \nabla^2 f(x)\Delta\right.$$

$$= \left\| \int_0^1 \left[\nabla^2 f(x + t\Delta x_{nt}) - \nabla^2 f(x)\right] \Delta x_{nt}\, dt \right\|_2$$

$$\overset{\text{Lipschitz}}{\le} \frac{L}{2}\|\Delta x_{nt}\|_2^2 = \frac{L}{2}\|\nabla^2 f(x)^{-1}\nabla f(x)\|_2^2$$

$$\overset{\text{Strong convexity}}{\le} \frac{L}{2m^2}\|\nabla f(x)\|_2^2$$

Lecture 20:

+) More detailed convergence proof:

- Damped Newton phase:

By strong convexity $\nabla^2 f(x) \preceq MI \quad \forall x \in S$:

$$f(x + t\Delta x_{nt}) \leq f(x) + t\nabla f(x)^T \Delta x_{nt} + \frac{M}{2}\|\Delta x_{nt}\|_2^2 t^2$$

$$\leq f(x) - t\lambda(x)^2 + \frac{M}{2m} t^2 \lambda(x)^2 \leq f(x) - \frac{t}{2}\lambda(x$$

Noting step size $\hat{t} = \frac{m}{M}$ satisfies backtracking line search, then line search must return $t \geq \beta\frac{m}{M}$, thus

$$f(x^+) - f(x) \leq -\frac{t}{2}\lambda(x)^2 \leq -\alpha t \lambda(x)^2$$

$$\leq -\frac{\alpha\beta m}{M^2}\|\nabla f(x)\|_2^2 \leq -\alpha\beta\frac{m}{M^2}\eta^2$$

[Here we use $\lambda(x)^2 = \nabla f(x)^T \nabla^2 f(x)^{-1} \nabla f(x)^T \geq \frac{1}{M}\|\nabla f(x)\|_2^2$

Thus choose $\gamma = \alpha\beta\frac{m}{M^2}\eta^2$ satisfies the damped phase.

- Quadratic phase:

For this phase we can always use step size $t = 1$ in backtracking search. We skip the proof for this here.

Now $x^+ = x + \Delta x_{nt}$ and

$$\|\nabla f(x^+)\|_2 = \|\nabla f(x + \Delta x_{nt}) - \nabla f(x) - \nabla^2 f(x)\Delta x_{nt}\|_2$$

$$= \left\| \int_0^1 \nabla^2 f(x + t\Delta x_{nt}) \cdot \Delta x_{nt}\, dt - \nabla^2 f(x)\Delta\right.$$

$$= \left\| \int_0^1 \left[\nabla^2 f(x + t\Delta x_{nt}) - \nabla^2 f(x)\right]\Delta x_{nt}\, dt \right\|_2$$

$$\overset{\text{Lipschitz}}{\leq} \frac{L}{2}\|\Delta x_{nt}\|_2^2 = \frac{L}{2}\|\nabla^2 f(x)^{-1}\nabla f(x)\|_2^2$$

$$\overset{\text{strong convexity}}{\leq} \frac{L}{2m^2}\|\nabla f(x)\|_2^2.$$

Can show $\eta = \min\{1, 3(1-2\alpha)\}\frac{m^2}{L}$.

Summary for Newton's method:

- Convergence rapid, quadratic near x^*.
- Affine invariant, thus insensitive to choice of coordinate or condition number of sublevel sets.
- The condition number of sublevel sets (or the Hessian) only affects the numerical inversion of the Hessian, but has little effect on the rate of convergence. Newton's method can tolerate large condition numbers of up to 10^{10}, whereas gradient descent tolerates far smaller number (pratically useless for $K > 1000$).
- Newton's method scales well with problem size. For example, for problems in $\mathbb{R}^{10}$ and $\mathbb{R}^{1000}$ the number of iterations can be comparable.

$\ominus$ The difficulty of the Newton's method is the cost of forming (computing) and storing the Hessian, and the cost of computing the Newton step, requires to solve $\nabla^2 f(x)\, \Delta x = -\nabla f(x)$.

Roughly the cost of computing the inverse of an $n \times n$ Hessian matrix is of order n^3. But for some problem we can exploit the structure of the Hessian to reduce this cost.

+) Some variants of Newton's method:

The goal is to reduce the computational complexity of Newton's.

- Quasi-Newton methods: Replace $\nabla^2 f(x)$ by approximation H. Many update rules for H, all satisfies:
  - $H = H^T \succ 0$
  - secant condition: $\nabla f(x^+) - \nabla f(x) = H^+ (x^+ - x)$
  - $H^{-1} \nabla f(x)$ is more easily computed than $\nabla^2 f(x)^{-1} \nabla f(x)$

Broyden - Fletcher - Goldfarb - Shanno (BFGS): (most common)

$$y = \nabla f(x^+) - \nabla f(x) \quad , \quad s = x^+ - x$$

then
$$H^+ = H + \frac{y y^T}{y^T s} - \frac{H s s^T H}{s^T H s} \rightarrow \mathcal{O}(n^2)$$

7) Self-concordance:

- Motivation:
  - Newton's method is affine invariant, but the convergen[ce] analysis is not
  - Often do not know constants m, M, L in practice
  - Constants m, M, L can depend on starting point.

- Self-concordance condition (Nesterov & Nemirovski) allows a new analysis of Newton's method
  - is affine invariant
  - involves no unknown constants
  - is valid for many function f, including the logarithmic barrier function (more on this later in interior-point metho[ds]

- Self-concordance condition:
  Convex $f: \mathbb{R} \to \mathbb{R}$ is self-concordance if

$$|f'''(x)| \leq 2 f''(x)^{3/2} \quad \forall x \in \mathrm{dom} f.$$

Examples:

- $f(x) = -\log x$
- $f(x) = x\log x - \log x$
- $f(X) = -\log\det X$

are SC.

- Some simple properties:
  - affine invariant:

$$f(x) \text{ SC} \iff g(z) = f(Az+b) \text{ is SC}.$$

  - Sum and scaling:

$$f, \hat{f} \text{ SC} \rightarrow f + \hat{f} \text{ SC}$$

$$f \text{ SC} \rightarrow \alpha f \text{ SC}.$$

Thus: $-\sum_i \log(b_i - a_i^T x)$ is SC

$-\log\det(F_0 + x_1 F_1 + \ldots + x_n F_n)$ is SC.

+ Analysis of Newton's method for SC functions:
Can show that with backtracking or exact line search:

$$\#\text{ iterations} \leq \frac{f(x^{(0)}) - f^*}{\eta_2} + \log_2 \log_2\left(\frac{2}{\varepsilon}\right)$$

where η_2 depends only on backtracking linesearch parameters:

$$\eta_2 = \beta \frac{\alpha(1/2 - \alpha)^2}{5 - 2\alpha}.$$

+ SC functions allow a more explicit analysis of convergence and complexity.
It is not known if SC functions are easier to minimize than non-SC functions.